新聞記事におけるトピック検出と局面追跡に関する研究

新聞記事におけるトピック検出と局面追跡に関する研究

A STUDY ON TOPIC DETECTION AND ASPECTS TRACKING IN NEWS ARTICLES

井上 眞乙
Maoto INOUE

指導教員 三浦 孝夫

法政大学大学院理工学研究科システム理工学専攻修士課程

In this investigation, we discuss aspect tracking, i.e.,how to identify tracking storylines of document topics. Since there happen huge amount of fragment information, it is hard to see what they mean and how they go within topics by hands. Here we attack to this kind of problems by means of stochastic models. Our basic idea is that we consider state transitions as internal structure of stories based on HMM, and we extract several storylines as aspects of topics by probabilistic likelihood. We utilize KL divergence to screen topics.
Key Words *: Aspect Tracking, Hidden Markov Model (HMM), Topic Detection and Tracking*

1. 問題の背景

現在, インターネットの普及により膨大な量の情報が日々増加しているが, 中には断片的な情報も多く存在し, 時間の経過と共にその内容が変化していく. 情報の大部分は文書であり, どの語がどの文書に含まれているかを調べる必要がある.

トピックに関する代表的なアプローチの 1 つとして,Topic Detection and Tracking(TDT) がある.TDT はある時点で発生する固有の事象を検出し, 状況の発展を追跡する事を可能にする. 一般的な情報源はラジオニュースや新聞記事である.TDT における New Event Detection は, 新規又は以前に未確認の事象に関連する文書を特定することであり, 主な技術は教師無クラスタリングである.Event Tracking は, 特定のトピックに対するストーリーを追跡する.Allan らは適合性フィードバックなど 2 つのアプローチを適用し, 情報フィルタリングに類似した追跡手法を提案している.Yang らは決定木 (dtree) と k 近傍法 (k-NN) を用いた手法を提案している. しかしこれらは事前にある程度の学習データを必要としている.

本研究では HMM を用いた尤度によるトピックの局面追跡を行う手法と, 語の増分を考慮した確率的 TF-IDF を用いたクラスタリングによるトピック検出を行う手法を提案する. これにより正確なオンラインでの検出・追跡のモデル化を実現する. 本研究の貢献は以下の 3 つである.

(1) 文書トピックの局面を追跡し, 内容の変遷を正確に捉える.

(2) 正確なオンライン検出のモデル化を可能にする確率的 TF-IDF を用いたトピック検出を議論する.

(3) 各推定に打ち切り閾値を設け終了の判断をし, 計算量の削減を図る.

2. 扱う問題

(1) トピック検出と局面追跡

TDT においてイベントは特定の時間と場所で発生する事象として定義され, トピックはある一連の事件を表す特定の話題のことである. 例えば,「千葉・市川女性刺殺事件」と「山口・周南放殺人事件」では「事件が発生した」「犯人逮捕」などの記事を含むトピックだが, それぞれの「殺人事件」は異なるストーリーを含んでいるため, 別々の事件を表す特定の話題として捉える方が自然である.

トピック検出は同様のトピックについて議論するイベント, 話題の変化を検出する技術である. 主に 2 つのアプローチで行う. 全ての記事が用意された上で行う遡及的検出およびストリーム内の記事を順に処理するオンライン検出である. ストーリーラインは, 一貫した意味のある事柄で記述される幾つかの適切な構造に言及する 1 つまたは複数のトピックに関連するイベントの系列である. 言い換えれば, 関連するイベントはストーリーの局面を構成する.

局面追跡と呼ばれるストーリー抽出 (新聞記事から) は, 特定のトピックの下でイベントを抽出し, これらのイベントが時間の経過と共にどのように発展するかを明らかにすることを目的としている. しかしここでのトピック追跡は, 単語の特徴などの共通項を基に議論しているのであり, 本研究が目指す局面追跡の” 前述・後述の記事の整合性に重きを置いたトピックの抽出および知識構造の発見” とは目的が異なる. 関連するイベントを正確に抽出し, それらを首尾一貫したストーリーに繋げるための技術が必要である.

局面追跡には主に 2 つの問題がある. 第 1 に, ストーリーの決まった構造がないため学習データを事前に用意する方法が無い. 同ジャンルのトピックであればストーリーの変遷が似てしまうため, 特定のトピックに対するストーリーの捕捉が難しい. 第 2 に, 急激な変化に対する対応が難しく, 同じトピックでありながら「異なるトピックの記事である」と間違った追跡を行う場合がある. また, トピック検出にも急激に増加する語分布の変化への対応が困難な問題があり, 語の増分を考慮したオンラインでの検出方法が必要となる.

本研究は追跡問題に対し,HMM による確率論的アプローチをとる. 状態遷移をストーリーとして考え, マルコフモデルを構築し, トピックの構造の可視化, 尤度計算による局面

追跡を行う. 尤度計算では Viterbi アルゴリズムを拡張した K-Viterbi アルゴリズム, 後ろ向き K-Viterbi アルゴリズムを使用し, 尤度上位 K 個の経路を得る. また, 検出においては増分を考慮した確率的 TFIDF および遡及性を考慮したクラスタリングによるオンラインでの検出方法を提案する.

(2) 提案手法

本研究では文書とする新聞記事に対し,HMM によるトピックの局面追跡および確率的 TF-IDF によるオンラインでのトピック検出の新規のアプローチを提案し, それぞれの手法を補完するように組み込んで行う. 局面追跡は, ある記事が特定のトピックにおける後続の記事となるかを 1 つのモデルで判断するため, クラスごとにモデルを構築して最尤原理 (MLP) により分類する分類問題とは異なる点に注意されたい.

トピック検出において, 新しい文書が継続的に入る事で文書数が変化し続ける状況を考慮すると, 毎回全ての文書の TF-IDF を計算するには時間がかかるため, 確率的逆文書頻度 (IDF) を適用する. また, トピックの局面追跡において, トピックの構造を確率過程の状態遷移とみなし, モデルの構築および尤度計算により類似する記事を選定することで構造抽出が可能と考える.Viterbi アルゴリズムを拡張した 2 つの方法,K-Viterbi アルゴリズム, 後ろ向き K-Viterbi アルゴリズムを提案し, 尤度計算を行う. 時系列に並べた記事系列を一定区間で区切り, 以下の手順に従い区画ごとに検出と追跡を行う. なお, 第 1 区画には追跡したいトピックの記事が数記事含まれているものとする.

(Step A) 現区画の最後の記事までトピック検出を行う. その後, 得られた各クラスタ i 内外の纏まりの良さ C_{CD_i} を中心度 C_{C_i} および密度 C_{D_i} に基づいてそれぞれ求め, 閾値 α と比較する.

$$
\begin{aligned}
C_{D_i} &= \frac{W_d}{W_{C_i}} \\
C_{C_i} &= \frac{1}{C_{N-1}} \sum \frac{(W_{C_{ij}})^2}{W_{C_i} \times W_{C_j}} \\
C_{CD_i} &= \frac{1-C_{D_i}+C_{C_i}}{2}
\end{aligned}
$$

(Step B) 閾値 α 以上であれば検出を完全終了し, 以下であれば検出は次の区画でも続行する. どちらの場合においても, 得られた学習データのクラスタを用いて, 現区画内のその他の記事 (テストデータ) の局面追跡を行う.

(Step C) 追跡出来た記事がある場合, その時点でモデルの性能 $M = D + P + \sum_D PP$ を求め, 閾値 β と比較する. 閾値以上であれば追跡を完全終了し, 以下であれば次の区画でも続行する.

(Step D) 現区画内の追跡が終わり次第, 次の区画と合体し Step A〜Step D を繰り返す. 検出または追跡が終了している場合はその推定は行わない.

クラスタ内外の纏まりの良さにおいて,W_d はクラスタ C_i 内で 2 件以上の記事 d に出現する語数 W,W_{C_i} はクラスタ C_i 内の全語数である.$W_{C_{ij}}$ はクラスタ間で共通する語数であり,C_{N-1} は比較したクラスタ数である. モデルの性能において,D はデータ数,P は推定されたパラメタ数であり,$\sum_D PP$ はパープレキシティ値の合計である.

3. 結果

各手法における各トピックの検出および局面追跡の結果を表 1, 表 2 にそれぞれ示す.

表 1 各手法における性能評価値 (トピック検出)

富山		ベースライン (CMU 最適化)	ベースライン (CMU)	提案手法
	未検出率 (%)	0.7059 (12/17)	0.2941 (5/17)	0.2353 (4/17)
	誤検出率 (%)	0 (0/23)	0	0
	Cost	0.3	0.125	0.1
山口		ベースライン (CMU 最適化)	ベースライン (CMU)	提案手法
	未検出率 (%)	0.75 (9/12)	0.5833 (7/12)	0.0833 (1/12)
	誤検出率 (%)	0	0	0.0357 (1/28)
	Cost	0.225	0.175	0.1083
千葉		ベースライン (CMU 最適化)	ベースライン (CMU)	提案手法
	未検出率 (%)	0.4545 (5/11)	0.4545 (5/11)	0.6364 (7/11)
	誤検出率 (%)	0	0	0
	Cost	0.125	0.125	0.175

表 2 各手法における性能評価値 (局面追跡)

富山		ベースライン	提案手法
	未追跡率 (%)	0 (0/15)	0.75 (3/4)
	誤追跡率 (%)	0.4211 (8/19)	0 (0/23)
	Cost	0.5333	0.1111
山口		ベースライン	提案手法
	未追跡率 (%)	0.1 (1/10)	0 (0/1)
	誤追跡率 (%)	0 (0/24)	0 (0/28)
	Cost	0.0294	0
千葉		ベースライン	提案手法
	未追跡率 (%)	0.6667 (6/9)	0.8571 (6/7)
	誤追跡率 (%)	0 (0/25)	0 (0/29)
	Cost	0.1765	0.1667

検出および追跡の精度が向上している.

4. 結論

本研究では, 尤度計算による局面追跡と, 確率的 TF-IDF による遡及性を考慮したトピック検出の手法を提案した. 結果の一部として, 富山トピックの DCF は最大で 0.2,TCF は 0.42 以上も改善された. また, 千葉トピックでは検出と追跡のそれぞれの結果でお互いを補完し, 検出漏れを防いだ. 各手法で打ち切り閾値を設けた事により後の他トピックの記事を処理することなく終了し, 計算量の削減が可能となった.

参考文献

1) Inoue, M. and Miura, T.: Stochastic Approach to Aspect Tracking, 23rd International Conference on Natural Language & Information Systems (NLDB 2018), 2018.
2) 井上 眞乙, 三浦 孝夫: 文書トピックの局面追跡，第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM フォーラム 2018) , 2018.
3) Inoue, M. Shirai M. and Miura, T.: Sequence Classification based on
Active Learning, 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2017), 2017.
4) Inoue, M. and Miura, T.: Learning Actively for Sequence Classification, The 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2017), 2017.